宠物知识图谱的半自动化构建方法 *

袁　琦，刘　渊，谢振平，陆　菁

(江南大学 数字媒体学院, 江苏 无锡 214122)

摘　要：提出一种宠物知识图谱的构建框架。通过自顶向下的方式设计并构建了 Schema(概念)层，从半结构化和非结构化数据中进行知识抽取构建了数据层。在对非结构化数据的实体抽取方面，提出了一种条件随机场（CRF）与宠物症状词典相结合的症状命名实体识别方法。该方法利用症状词典对文本进行识别，获取语义类别信息，CRF 结合语义信息实现对症状实体的识别抽取。实验结果表明了该方法的有效性。在知识表示方面，选用 OrientDB 数据库支持的属性图模型来表示。知识图谱采用 OrientDB 图数据库来完成知识的存储，并实例展示了构建的宠物知识图谱。

关键词：宠物知识图谱；症状术语词典；宠物症状命名实体识别；条件随机场；图数据库

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.05.0490

Semi-automated construction method of pet knowledge graph

Yuan Qi, Liu Yuan, Xie Zhenping, Lu Jing

(*School of Digital Media*, *Jiangnan University*, *Wuxi Jiangsu* 214122, *China*)

Abstract: This paper proposed a construction framework of pet knowledge graph. It designed and constructed the schema layer in a top-down manner and constructed the data layer by extracting knowledge from semi-structured and unstructured data. For entity extraction of unstructured data, this paper proposed a symptom-named entity recognition method which combined conditional random field (CRF) and pet symptom dictionary. The method used symptom dictionary to identify the text and obtain the semantic category information, and then combined CRF and the semantic information to identify symptom-named entities. The experimental results showed the effectiveness of the method. The attribute graph model supported by the OrientDB database was selected for knowledge representation. The knowledge graph used the OrientDB graph database for knowledge storage. In addition, examples were shown for the constructed pet knowledge graph.

Key words: pet knowledge graph; symptoms dictionary; pet symptom named entity recognition; conditional random field（CRF）; graph database

0　引言

随着经济社会的发展，宠物越来越多地出现在人们生活当中，家庭结构和人口结构的变化使得宠物进入了更多的家庭。据京东《2017 宠物消费趋势报告》的分析，目前中国宠物已经突破 1 亿只。互联网是人们很重要的获取宠物百科知识和宠物医疗知识的来源之一。大多数的宠物主人缺乏宠物知识，当他们需要了解这方面的知识的时候，大多的宠物主人主要是通过互联网上 Google 和百度之类的搜索引擎来获取知识。然而这会花费宠物主人很多时间来判断哪些内容包含了自己想要的信息，很多时候，用户想要获取进一步的知识，还需要自己再一次的阅读和筛选。这导致了信息检索的效率比较低下，用户会对搜索引擎返回的大量信息感到迷茫。因此人们对可以提交用自然语言表达的宠物相关问题，系统会返回相关又准确的答案的问答系统有着非常迫切的需求。目前基于知识库的问答聊天机器人有微软小冰、百度的度秘等。因此构建关于宠物知识库对实现智能问答有研究意义和应用价值。

目前国内外大型互联网公司纷纷推出知识图谱以改善服务质量，同时当今也涌现出了人类医学的知识图谱，并且发展迅速。但在宠物领域尚未出现成熟、专业的知识图谱。本文的主要工作包括：

a）宠物知识图谱 Schema(概念)层构建。根据需求，利用并且分析基于有宠网的疾病百科来定义宠物知识图谱 Schema 层。

b）信息抽取。实体抽取、实体属性关系抽取和语义关系的

收稿日期：2018-05-31；**修回日期：**2018-07-09　**基金项目：**国家自然科学基金资助项目（61672264）；国家科技支撑计划资助项目（2015BAH54F01）；江苏省研究生科研与实践创新计划项目（SJCX17_0505）

作者简介：袁琦（1993-），男，江苏扬州人，硕士研究生，主要研究 方向为社交网络、知识图谱（893182191@qq.com）；刘渊（1967-），男，江苏无锡人，教授，博导，主要研究方向为数字媒体技术、网络安全和网络仿真等；谢振平（1979-），男，江苏常州人，副教授，博士，主要研究方向为演化认知学习、知识网络、机器视觉等；陆菁（1983-），女，浙江金华人，讲师，主要研究方向为数字媒体艺术.

抽取。从不同数据源中通过爬虫爬取，数据过滤、清洗、解析来获取结构化宠物知识和实体属性关系抽取、语义关系的抽取。通过条件随机场（CRF）与症状字典相结合的症状命名实体识别模型来获取命名实体。首先通过爬取网上知识来构造宠物医学症状相关的术语及语义类别信息词典。通过将症状的语义类别信息作为特征加入到 CRF 模型中来获取比较准确的疾病症状命名实体识别。

c）知识表示。选择 OrientDB[1]原生图数据库支持的属性图模型来进行知识表示。

d）将获取到的 Schema 层数据和实例层数据通过 OrientDB 图数据库进行知识的存储, OrientDB 图数据库使用类 SQL 查询语句。

1　相关工作

2012 年，谷歌正式提出知识图谱(knowledge graph) [2]的概念，以此为基础构建智能搜索系统，希望通过准确了解用户的搜索意图，改善搜索质量和用户的搜索体验。

目前很多国外的大规模通用知识图谱的研究已经有了很多成果，具有代表性有 YAGO[3]、Freebase[4]、DBpedia[5]、NELL[6]等，这些知识图谱包含了大量的常识知识。与国外相比，国内知识图谱构建与研究还处于起步阶段，主要有 Zhishi.me[7]等。

国内垂直领域知识图谱，如中医药知识图谱的构建[8]，根据领域知识创建中医药知识图谱的模式，通过信息转换，将关系数据库中的中医药结构化信息转换为 RDF 数据，在信息抽取模块，采用多策略学习的方法，从半结构化和非结构化数据中抽取信息，最后将不同数据源的数据进行模式对齐。

文物知识图谱的构建[9]，通过在七步法和骨架法上进行改进，提出了文物的本体构建方法；接着设立知识节点，知识存储，通过采集过来的文物信息进行本体实例化，完成之后的文物实例和本体概念就是知识图谱中的知识节点；最后以三元组的形式存储到图数据库 Neo4j 中。

面向开源软件项目的软件知识图谱的构建[10]，针对四种不同类型的软件资源，提出了软件知识实体的提取原则和方法，提出了软件知识实体之间关联关系构建的方法，设计了软件知识图谱的构建框架，由软件知识提取模块、知识融合模块、存储管理和软件知识检索模块构成。

双语影视知识图谱的构建研究[11]，通过半自动化的方法构建了双语影视本体，在知识链接方面，采用基于 Word2Vec 和 TFIDF 两种实体相似度计算方法，在实体匹配上面，提出基于相似度传播算法的实体匹配算法。总的来说，国内垂直领域的知识图谱还是比较少，在宠物领域方面，目前国内还没有高质量的宠物知识图谱。

大数据环境下历史人物知识图谱构建与实现[12]，将数据以结构化的方式呈现，建立以关键词为中心的知识体系，通过采用基于标签遍历以及基于链接权重的方法进行数据的解析，之后将获取到的数据存储到历史人物库，在知识图谱的基础上进行可视化。

通常领域知识图谱注重知识的层次结构，需要预先构建模式图（schema 层）。本文采用的是大部分本体知识库都采用的半自动化的知识图谱构建方法。通过自顶向下的方式构建模式图（schema 层），也就是通过手工方式先构建宠物知识图谱的概念层，通过自底向上的方式构建宠物知识图谱的数据图（数据层），利用多种抽取技术获得实体、属性以及关系，抽取出高置信度的知识合并到知识图谱。

在构建知识图谱的过程中，需要对描述症状的非结构化文本数据进行症状的命名实体识别。目前有很多常用的机器学习模型用来解决命名实体识别问题，如隐马尔可夫模型（HMM）[13]、支持向量机（SVM）[14]、条件随机场（CRF）等。

CRF 是由 Lafferty 等人[15]于 2001 年在隐马尔可夫模型和最大熵模型的基础上提出的统计序列标注算法。CRF 可有效地克服隐马尔可夫模型（HMM）假设条件的限制以及能在一定程度上解决标记偏置问题。CRF 可以看做是一种无向图模型。常用的 CRF 模型是线性链 CRF。给定输入句子中的单词序列作为观测序列 o，s 表示对应的输出标记序列，CRF 定义了 s 的条件概率分布 $p(s|o)$,通过训练求得 $p(s|o)$ 为最大值时的状态序列 s。线性链 CRF 中的输出序列 s 的条件概率公式入下：

$$p(s|o) = \frac{1}{z}\exp(\sum_i\sum_k \lambda_k f_k(s_{i-1}, s_i, o, i)) \tag{1}$$

$$z = \sum_s \exp(\sum_{i=1}^{I}\sum_k \lambda_k f_k(s_{i-1}, s_i, o, i)) \tag{2}$$

其中：s 为标注序列；o 为观察序列；z 是归一化因子，为的是使状态序列 s 的概率和为 1；$f_k(s_{i-1}, s_i, o, i)$ 为特征函数；λ_k 是对应的权值，通常采用 L-BFGS 算法对条件随机场 CRF 进行参数估计。CRF 模型在近几年被广泛地运用到医疗命名实体识别当中，并取得了不错的效果。

2　宠物知识图谱的构建

宠物知识图谱设计并构建的总体框架包括五个步骤，如图 1 所示。

a) Schema 层的构建。采用自顶向下的方式构建宠物知识图谱的概念层。

b) 从半结构化数据中抽取。从半结构化的数据源中进行实体、关系和属性的抽取。

c) 从非结构化数据中抽取。从非结构化的数据中进行命名实体识别和抽取。

d) 知识表示。宠物知识图谱使用的是属性图模型来进行知识表示。

e) 知识存储。宠物知识图谱使用了 OrientDB 图数据库存储引擎存储获取到的宠物知识数据。

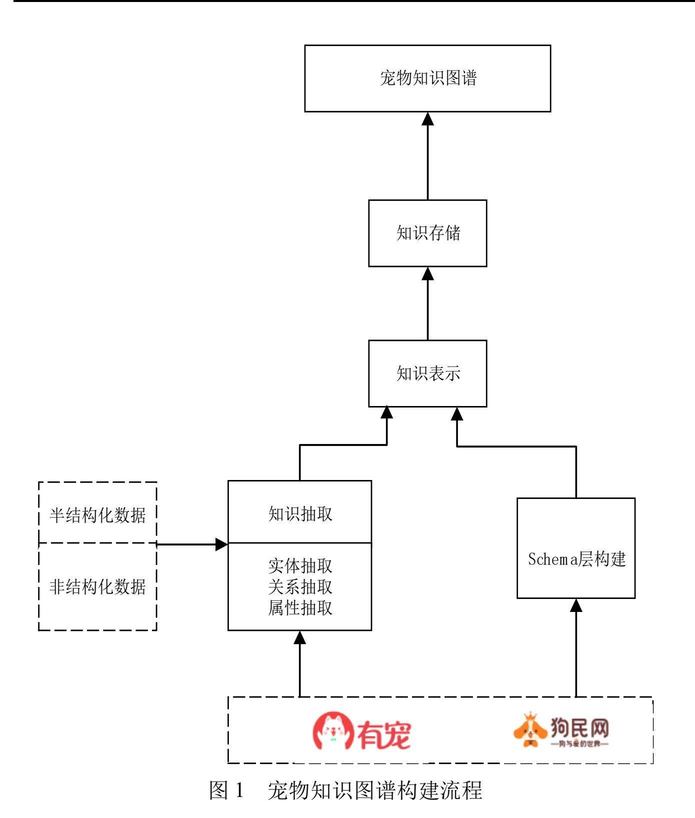


图 1　宠物知识图谱构建流程

Fig.1　Pet knowledge graph construction process

2.1　Schema 层的设计与构建

Schema 层的构建是对整个宠物知识图谱框架的构建，Schema 是要对类及类之间的关系进行定义，也就是对知识图谱中的概念与概念之间的语义关系进行定义。

本文构建的是宠物领域的知识图谱（宠物狗和猫为主），设计并构建了宠物领域知识图谱的 Schema 层，定义了基本的四大类，其中包括四大类宠物品种、宠物疾病、疾病症状和宠物食物。

其次是属性的定义：

a）宠物品种的属性包括中文名、别名、体型、毛长、英文名、智商、原产地、体重、寿命、价格、肩高、毛色和功能。

b）宠物疾病的属性定义包括科属、概述、发病原因、诊断标准、治疗方法和防治方法。

c）宠物食物的属性定义包括可食性。

以上是经过分析的宠物品种、宠物疾病和宠物食物的属性，疾病症状比较特殊，只有症状名称，不存在属性关系的定义。

根据定义的四大类，创建了三种语义关系，分别是：

a）e_HasDisease(有疾病)。宠物品种——宠物疾病，宠物品种与宠物疾病之间存在关系。

b）e_HasSymptom(有症状)。宠物疾病——疾病症状，宠物疾病与疾病症状之间存在关系。

c）e_EatFood(吃食物)。宠物品种——宠物食物，宠物品种与宠物食物之间存在关系

以上就是宠物知识图谱概念与语义关系的创建。宠物知识图谱的 Schema 如图 2 所示。

2.2　数据源

宠物知识图谱是从国内的关于宠物的网站上抽取知识。本文主要从“铃铛宠物”和“有宠”两家宠物网站爬取有用的知识，其中在铃铛宠物抽取了 92 种食物的实体，以及食物的属性。有宠网站上面有关于宠物品种和宠物的疾病的百科知识，提供了质量较高的半结构化数据，所以在有宠网站抽取了 1 367 个实体。

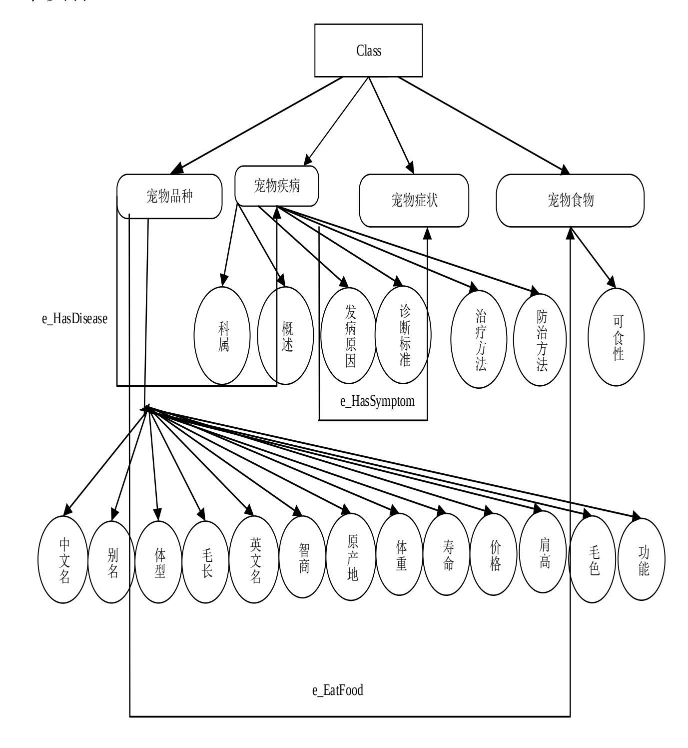


图 2　宠物知识图谱 Schema 层

Fig.2　Pet knowledge graph Schema layer

2.3　从半结构化数据中抽取

本文主要从“铃铛宠物”和“有宠”这两个网站的半结构化数据中抽取宠物品种、宠物疾病和宠物食物的实体、实体属性以及语义关系。采用的是网页爬虫和数据解析。通过爬虫采集网页信息。

本文选用可以从 HTML 网页中提取数据的 python 库—Beautiful Soup 作为解析器。基于网页页面布局相似的特点，采用基于标签遍历的方法，直接导航到 DOM 树的关键节点，可以避免大量遍历节点，从而提取相关网页正文。通过此方法，可以抽取宠物品种以及属性、宠物疾病以及属性、宠物食物以及食物属性的实体。同时在抽取实体的过程中也实现了语义关系的挖掘，获取了三种语义关系。宠物狗的阿司匹林中毒疾病如图 3 所示。

图 3 中，解析网页抽取了宠物疾病实例宠物狗的阿司匹林中毒，也抽取了阿司匹林中毒的科属、概述、发病原因、诊断标准、治疗方法五个属性，根据本文对宠物疾病属性的定义，也就获取了五条“属性—值”关系，用三元组描述为<实体，属性，属性值>，同时这是宠物狗的疾病，也就获取了 e_Hasdisease（有疾病）这种语义，图 3 中基本资料里面的症状不全也不正确，还需要从主要症状中抽取，主要症状是一段非结构化文本.本文将采用 CRF 与症状词典相结合的症状命名实体识别方法进行症状实体的抽取，这可以得到 e_Hassymptom(有症状)这种

语义。

阿司匹林中毒

基本资料
　科属：中毒
　症状：食欲下降口吐

概述
　抑制前列腺素合成，大剂量阿司匹林能阻止氧化磷酸化的过程，但也可能会引起高血糖。病猫患病初期兴奋呼吸，后期则转为抑制呼吸。会出现代谢性酸中毒、血小板凝集作用下降、骨髓发育受阻症状。

发病原因
　意外吞食阿司匹林(乙酰水杨酸)或药物使用剂量不当。幼犬由于体内缺乏代谢酶，尤其是缺乏合成葡萄糖醛酸化物的酶，而易发本病。病犬一次服用量>60m g/kg，可引起中毒。

主要症状
　中毒早期出现呼吸急促，后期则抑制呼吸；会出现体温升高、食欲下降、呕吐、溃疡性肠炎、代谢性酸中毒等症状，严重时出现昏迷、肾脏功能受损、出血等症状，长期用药会引起病犬非再生性贫血；偶见抽搐。

诊断标准
　诊断了解病史对本病的诊断十分有益；代谢性酸中毒、尿酸、阴离子间隙增大；血清或尿中的水杨酸含量具有一定诊断意义，取lm l尿液，酸化后加入3滴10%氯化铁，出现红色，表明水杨酸阳性；应与其他引起胃炎及严重代谢性酸中毒的疾病，乙二醇中毒，其他非类固醇抗菌消炎药物，如布洛芬中毒区别。

治疗方法
　动物摄入阿司匹林应尽早催吐、洗胃、服用活性炭及导泻药物，阻止毒物进一步吸收；碱化尿液36～48h，促进毒物的排出：碳酸氢钠，50m g/kg，口服，每天2～3次；碳酸氢钠也可缓解机体的代谢性酸中毒。支持疗法：补液、补充电解质，维持酸碱平衡；应用胃肠道保护剂及组胺受体颉颃剂(甲腈咪胍、甲胺呋硫)；病情严重者进行碱性腹膜透析液析。

图 3　阿司匹林中毒疾病

Fig.3　Aspirin poisoning disease

2.4　从非结构化数据中抽取

本文需要从非结构文本中进行命名实体识别来抽取症状的实体。在目前现有的很多机器学习算法中，CRF 不仅可以使用包括字、词、词性在内的多种上下文特征，还可以结合词典等外部特征，在命名实体识别等任务中取得了较好的效果。本文因此研究采用了 CRF 与症状词典结合的方法。症状命名实体识别的关键技术框架如图 4 所示。

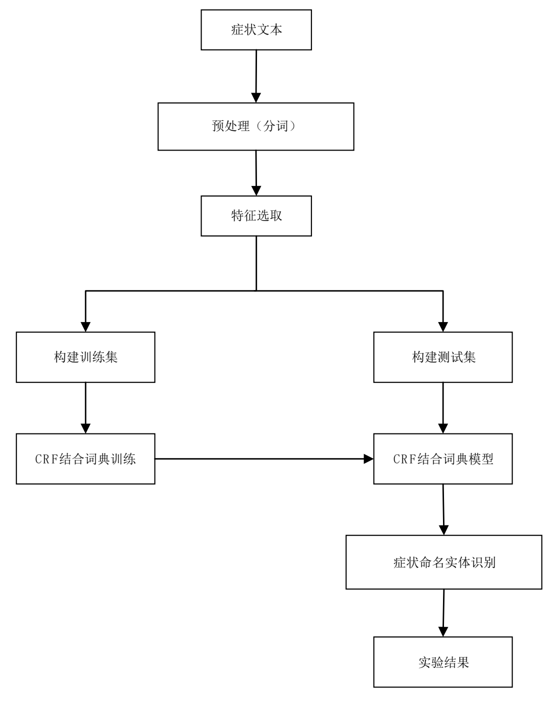


图 4　症状命名实体识别关键技术框架

Fig.4　Key technical framework for symptomatic named entity recognition

2.4.1 数据集与标注

经查阅文献和网上资源后，发现国内外目前没有公开的宠物医疗领域的用于症状命名实体识别的数据集，因此本文需要自己构造语料库。本文共抽取了 285 条描述症状的文本，其中将 100 条构建成训练集，30 条文本构建成测试集。当准确率达到要求时，将用训练好的模型来从 285 条无结构化文本中抽取症状的实体。

标记完语料库之后，需要对语料库进行格式转换，按照 BIESO 对语料进行标志。标记为 B-SIGNS、I-SIGNS、E-SIGNS、S、O，分别标志症状的首部、症状的中部、症状的尾部、单个症状词和非症状词。表 1 为使用 BIESO 标记实体的举例。

表 1　BIESO 标记实体举例

Table 1　Examples of BIESO marked entities

句子	BIESO 标记
病犬鼻腔粘膜呈现	病犬/O 鼻腔/B-SIGN 粘膜/I-SIGNS 呈现/I-SIGNS
潮红、肿胀	潮红/E-SIGNS、/O 肿胀/S-SIGNS

2.4.2 CRF 与症状词典相结合的命名实体识别方法

由于需要从描述症状的非结构化文本中抽取症状实体，所以采用了 CRF 与症状词典相结合命名实体识别方法。主要是通过网上查找分析，构造一个症状的词典，这样就可以利用症状词典获取文本中词语的语义类别信息，并把语义类别信息作为特征传递给 CRF 模型去识别文本中的症状实体。类别信息如表 2 所示。本文将描述症状的文本分为两类，即描述症状的术语记为“BS”，其他非症状术语记为“BO”。

表 2　类别信息

Table 2　Category information

类别	描述	举例	标记
症状术语	宠物因疾病而导致的异常表现或不适	呕吐、呼吸急促	BS
其他	文本中的其他词汇	病犬、长期用药	BO

2.4.3 特征选取

特征集是症状实体识别成功的关键。为了提高命名实体识别的准确率，通过对描述症状的文本分析。本文特征集包括“word”语言符号特征、词性特征以及症状词典特征，如表 3 所示。

表 3　症状特征

Table 3　Symptom characteristics

序号	特征	描述
1	Word	当前词的字符信息
2	Pos	当前词的词性
3	dict	当前词在症状术语中的语义类别

1）“word”语言符号特征　Word 语言符号特征指的是词的本身，包含丰富的有效信息。词是一种语言符号，本身可以作为一种特征，反映字符信息。与英文不同，中文之间没有明显的空格分隔符，所以在进行症状的实体识别之前，需要将文本进行分词，之后将分词结果作为 word 特征引入。

2）“pos”词性特征　在宠物疾病症状的实体识别任务中，

文本中的症状实体通常出现在动词后面，所以将词性作为特征，主要包括动词、名词、副词等。

3）"dict"词典特征 文本中包含大量专业症状名词，所以需要引入词典特征，通过构造的症状术语词典，用该词典匹配文本单词，结果返回症状的语义类别。词典特征就是症状词典对当前单词的识别结果，分为"BS"和"BO"。

2.4.4 实验与结果

本文总共有 285 条非结构化文本数据，其中使用标注的 130 条数据集做实验，将 100 条文本作为训练集，30 条作为测试集来进行实验。为了得到可靠稳定的模型，采用基于训练集的 10 折交叉验证，从而得到 CRF 模型的最优参数，并于单独的 30 条测试集上测试。实验采用的是机器学习常用的评价指标 precision(准确率)、recall（召回率）以及 F 值（F-measure）。具体定义如下：

$$P = \frac{\text{正确识别出的实体个数}}{\text{识别出的实体个数}} \times 100\% \qquad (3)$$

$$R = \frac{\text{正确识别出的实体个数}}{\text{标准结果中的实体个数}} \times 100\% \qquad (4)$$

$$\mathrm{F} = \frac{2PR}{P+R} \times 100\% \qquad (5)$$

进行对比实验的硬件平台为戴尔 Alienware Aurora R7，CPU 3.7 GHz Intel Core i7，RAM 32 GB，硬盘 2 TB+512 GB SSD。分为加词典特征和不加词典特征的对比实验，进行两个实验来看识别症状实体的实验效果。实验结果如表 4 所示。

表 4 实验结果对比

Table 4 Comparison of experimental results

方法	precision	recall	F-measure
CRF	0.8413	0.8172	0.8291
CRF+dict	0.8978	0.8817	0.8897

通过对比实验，结果显示结合了动物症状术语词典的 CRF 模型比没有结合词典特征的 CRF 模型识别效果有了不错的提升，准确率、召回率和 F 值都提高了不少，分别提升了 6.71%、9.08%和 7.90%，其中召回率提升幅度最大。分析实验结果发现，症状识别效果的提升原因是因为在症状的描述训练集中很少出现的，不具有明显特征的症状被结合症状词典的 CRF 模型准确地识别了出来，如"多饮多尿"，在本文的训练集中没有这样描述症状的术语，但是动物症状术语词典识别了出来，结合动物症状词典的 CRF 模型，因为有了语义类别信息，识别效果比未结合动物症状术语的 CRF 模型好。

因为识别出来的准确率达到了 91.63%，召回率达到了 90.32%，准确率和召回率都达到了比较高的数值，所以本文采用了这个训练好的结合症状词典的 CRF 模型从 285 条非结构化文本中抽取症状实体，共抽取出了 624 个宠物疾病的症状实体。

2.5 知识表示

知识图谱也可以看做是一种图的网络结构，网络图中的节点表示实体，边代表关系。知识图谱图模型可以使用 W3C 提出的资源描述框架（resource description frame，RDF）或者属性图（property graph）来表示[16]。本文因为使用 OrientDB 图数据库来存储获取到的宠物领域的数据，所以使用属性图模型来进行知识的表示。

属性图包含实体（节点）和链接实体的关系（边）。实体可以包含任何数量的属性（键值对形式），属性图中的元素如下：

a)一组节点。每个节点有唯一的标志符@rid，每个顶点有一组出边和入边，每个顶点都有个实体类型@class，表示实体所对应的概念类，每个顶点有键值对来定义属性集合。

b)一组边。每条边都有一个唯一标志符@rid，每条边有一个头节点和尾节点，每条边有个实体类型@class，表示两个节点之间的关系，每条边有键值对来定义属性结合。

图 5 描述了一个 OrientDB 的属性图模型，疾病"犬瘟热"实体和症状"发热"之间的关系是 e_Hassymptom(有症状)。其中@rid 是唯一标志符，@class 是实体类型，也就是对应的概念类，out 对应的是头节点也就是疾病节点，in 对应的是尾节点也就是症状节点，name 和 keshu 等键值对是对对应节点属性的描述。

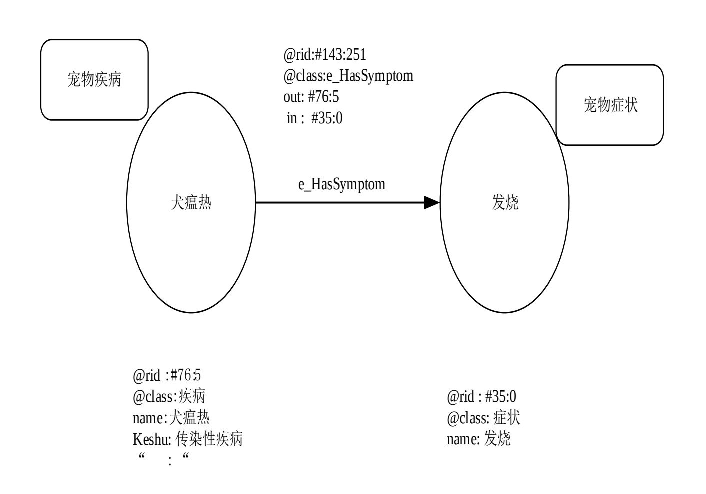


图 5 属性图实例

Fig.5 Property map example

2.6 知识存储

本文使用的是图数据库 OrientDB。OrientDB 是一个用 Java 实现的开源 NoSQL[17]数据库管理系统。它是一个多模式的库，支持图形、文档、键值对、对象模型和关系，也可以为图数据的管理与记录之间提供连接。支持的查询语言最常用的是 Gremlin[18]和 SQL[19]，用来操作属性图，支持以 SQL 的方式来查询数据，但是在标准的 SQL 上面扩展一些功能用来方便图的操作，是一种类 SQL 语句。

将获取到的宠物领域的实例层数据通过 OrientDB 原生数据库进行知识的整合和存储，存储语言使用类 SQL。首先需要创建模式，根据 Schema 层的定义，创建概念类包括宠物品种（v_Breed）、宠物疾病（v_Disease）、食物（v_Food）、疾病症状（v_Symptom）、有疾病（e_HasDisease）、吃食物（e_EatFood）

和有症状（e_HasSymptom）

在创建好模式之后，需要载入对应标签中的所有节点信息以及节点之间的关系。在导入数据信息时为了防止重复的节点信息和重复的关系，需要用类 SQL 查询语句进行判断。判断症状重复和载入症状信息的类 SQL 查询语句如下所示：

```
"LET $symptom = select from v_symptom WHERE name = '%s';"\
"if($symptom.size()<1){" \
    "CREATE VERTEX v_symptom SET name = '%s';" \
"}"%（symptom, symptom）
```

类 SQL 语句首先在图数据库中查询该症状实体，然后用到了 if 语句来判断症状实体是否已经存在，如果 symptom.size()小于 1 的话则表示该症状实体未在图数据库中出现,则要创建表示该症状的新的实体。

表 5 为全部数据存储到图数据库之后得到的相关详细信息。因为 OrientDB 内置集成了可视化工具，所以通过可视化可以看到“犬瘟热”这个疾病的所有症状的可视化结果，如图 6 所示。图中蓝色节点表示疾病犬瘟热；橙色节点表示犬瘟热的 9 个症状；边 e_HasSymptom 表示有症状。

表 5　整合后的知识库数据统计

Table 5　Integrated knowledge base data statistics

统计项	数值	统计项	数值
实体类型数目	4	疾病节点数目	285
关系类型	3	食物节点数目	92
属性数目	20	节点总数目	1358
症状节点数目	624	关系总数目	79527
宠物品种节点数目	357		

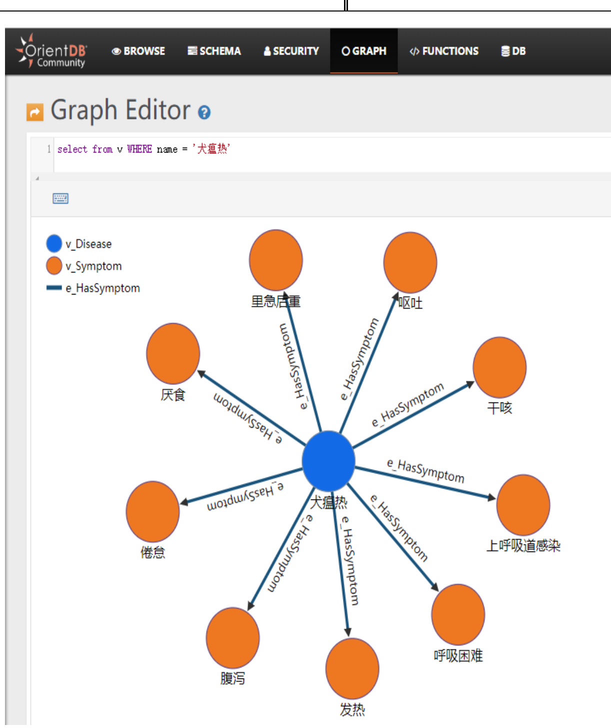


图 6　宠物知识图谱示例图

Fig.6　Example of pet knowledge graph

3　结束语

本文研究了在宠物领域一种基于数据抽取的知识图谱的构建方法，并且详细地描述了整个构建过程，通过实例展示了本文构建的知识图谱，旨在为宠物领域构建比较高质量的知识库。

首先采用自顶向下的方式构建了 schema（概念）层，对整个宠物知识图谱框架进行了构建，也就是对知识图谱中的概念和概念之间的语义关系进行定义；然后通过从半结构化的数据中进行实体、关系和属性的抽取，从非结构化数据中进行命名实体识别和抽取。在非结构化的知识抽取中，提出了 CRF 结合症状词典的命名实体识别方法来对症状实体进行识别获取，通过做实验表明，结合症状词典的 CRF 模型比未结合词典的 CRF 模型效果要好。获取完宠物知识之后，通过 OrientDB 数据库支持的属性图模型来进行知识表示。选用 OrientDB 原生图数据库来进行知识的存储，并且通过 OrientDB 内置的可视化，实例展示了构建的宠物知识图谱。

构建知识库是一项复杂性的工作，具有系统性和长期性。宠物知识图谱需要改进的地方还有很多，比如宠物领域知识还不够丰富，还需要寻求更多的宠物知识源来扩展知识库，并进行知识融合，包括实体对齐和模式对齐等。研究建立知识图谱的更新机制，在完善了宠物知识图谱之后，可以在此基础上进行智能问答的研究，用户使用自然语言提出问题，通过知识图谱的帮助下对问题进行语法分析和语义分析，将问题转换成结构化的查询语句，使用图的查询语句，在知识图谱中查询答案。

总的来说，本文设计并实现了基于数据抽取的宠物知识图谱，填补了国内在宠物领域知识图谱的缺失。该知识库为宠物领域知识的应用提供了语料基础，为宠物领域问答机器人奠定了基础，具有重要意义；同时本文提出的构建方法对垂直领域知识图谱的构建具有一定借鉴意义。

参考文献：

[1] Tesoriero C. Getting started with orientDB [M]. Birmingham: Packt Publishing Ltd, 2013.

[2] Pujara J, Miao H, Getoor L, *et al.* Knowledge graph identification [C]// Proc of International Semantic Web Conference. Berlin: Springer, 2013: 542-557.

[3] Hoffart J, Suchanek F M, Berberich K, *et al.* YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial In-Telligence, 2013, 194: 28-61.

[4] Yue Bin, Gui Min, Guo Jiahui, *et al.* An effective framework for question answering over freebase via reconstructing natural sequences [C]// Proc of International Conference on World Wide Web Companion; International World Wide Web Conferences Steering Committee. New York: ACM Press, 2017: 865-866.

[5] Ritze D, Bizer C. Matching Web tables to DBpedia-a feature utility study [J]. Context, 2017, 42 (41): 19.

[6] Santos F A O, do Nascimento F B, Santos M S, *et al.* Training neural tensor



networks with the never ending language learner [M]// Information Technology-New Generations. Cham: Springer, 2018: 19-23.

[7] Niu Xing, Sun Xinruo, Wang Haofen, *et al.* Zhishi. me-weaving chinese linking open data [C]// Proc of International Semantic Web Conference. Berlin: Springer, 2011: 205-220.

[8] 阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用 [J]. 医学信息学杂志, 2016, 37 (4): 8-13. (Ruan Tong, Sun Chenglin, Wang Haofen, *et al.* The construction and application of knowledge map of traditional Chinese medicine [J]. Journal of Medical Informatics, 2016, 37 (4): 8-13.)

[9] 林炀平. 文物知识图谱构建与检索关键技术研究与实现 [D]. 杭州: 浙江大学, 2017. (Lin Yangping. Research and implementation of key technologies for the construction and retrieval of cultural relics knowledge maps [D]. Hangzhou: Zhejiang University, 2017.)

[10] 李文鹏, 王建彬, 林泽琦, 等. 面向开源软件项目的软件知识图谱构建方法 [J]. 计算机科学与探索, 2017, 11 (6): 851-862. (Li Wenpeng, Wang Jianbin, Lin Zeqi, *et al.* Software knowledge mapping construction method for open source software projects [J]. Journal of Computer Science and Technology, 2017, 11 (6): 851-862.)

[11] 王巍巍, 王志刚, 潘亮铭, 等. 双语影视知识图谱的构建研究 [J]. 北京大学学报: 自然科学版, 2016, 52 (1): 25-34. (Wang Weiwei, Wang Zhigang, Pan Liangming, *et al.* Research on the construction of bilingual film and television knowledge map [J]. Journal of Peking University: Natural Science, 2016, 52 (1): 25-34.)

[12] 周亦, 周明全, 王学松, 等. 大数据环境下历史人物知识图谱构建与实现 [J]. 系统仿真学报, 2016, 28 (10): 2560-2566. (Zhou Yi, Zhou Mingquan, Wang Xuesong, *et al.* Construction and implementation of knowledge graph of historical figures in big data environment [J]. Journal of System Simulation, 2016, 28 (10): 2560-2566.)

[13] Leos-Barajas V, Photopoulou T, Langrock R, *et al.* Analysis of animal accelerometer data using hidden Markov models [J]. Methods in Ecology and Evolution, 2017, 8 (2): 161-173.

[14] Yan He, Ye Qiaolin, Zhang Tian'an, *et al.* Least squares twin bounded support vector machines based on L1-norm distance metric for classification [J]. Pattern Recognition, 2018, 74: 434-447.

[15] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th International Conference on Machine Learning. Burlington: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[16] 王昊奋. 大规模知识图谱技术 [J]. 中国计算机学会通讯, 2014, 10 (3): 64-68. (Wang Haofen. Large-scale knowledge graph technology [J]. Chinese Journal of Computer Science, 2014, 10 (3): 64-68.)

[17] Ahmadian M, Plochan F, Roessler Z, *et al.* Secure NoSQL: an approach for secure search of encrypted nosql databases in the public cloud [J]. International Journal of Information Management, 2017, 37 (2): 63-74.

[18] Thakkar H, Punjani D, Auer S, *et al.* Towards an integrated graph algebra for graph pattern matching with gremlin [C]// Proc of International Conference on Database and Expert Systems Applications. Cham: Springer, 2017: 81-91.

[19] Cattell R. Scalable SQL and NoSQL data stores [J]. ACM Sigmod Record, 2011, 39 (4): 12-27.